



(51) International Patent Classification:
G06F 13/00 (2006.01)

(21) International Application Number:
PCT/US2015/043210

(22) International Filing Date:
31 July 2015 (31.07.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/031,616 31 July 2014 (31.07.2014) US

(71) Applicant: CORNELL UNIVERSITY [US/US]; Center For Technology Licensing ("CTL") At Cornell University, 395 Pine Tree Road, Ithaca, NY 14850 (US).

(72) Inventors: PASS, Rafael; 849 Carroll St., #4, Brooklyn, NY 11215 (US). BOYLE, Elette; P.O. Box 23467, 6123401 Tel Aviv (IL). CHUNG, Kai-Min; 2f., No.1, Ln. 365, Sec. 1, Mingzhi Rd., Taishan Dist., New Taipei City, 24347 (TW).

(74) Agents: VALAUSKAS, Charles, C. et al.; Valauskas Corder LLC, 150 S. Wacker Drive, Suite 620, Chicago, IL 60606 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to the identity of the inventor (Rule 4.17(i))
- of inventorship (Rule 4.17(iv))

Published:

- with international search report (Art. 21(3))

(54) Title: OBLIVIOUS PARALLEL RANDOM ACCESS MACHINE SYSTEM AND METHODS

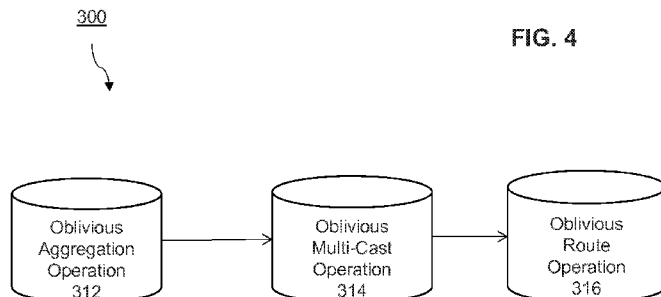


FIG. 4

(57) Abstract: A secure parallel processing technique for converting any parallel random access machine (PRAM) into an oblivious parallel random access machine (OPRAM), whose distribution of memory accesses is statistically independent of the data (with negligible error), while only inducing a polylogarithmic slowdown to both the total and parallel complexities of the program.



OBLIVIOUS PARALLEL RANDOM ACCESS MACHINE SYSTEM AND METHODS

CROSS-REFERENCE TO RELATED APPLICATIONS

5 This application claims the benefit of U.S. Provisional Patent Application No. 62/031,616 filed July 31, 2014, incorporated by reference.

STATEMENT REGARDING FEDERALLY-SPONSORED RESEARCH OR DEVELOPMENT

10 This invention was made with government support under AFOSR YIP Award FA9550-10-1-0093, and DARPA and AFRL under contract FA8750-11-2-0211. The U.S. government has certain rights in the invention.

FIELD OF THE INVENTION

15 The invention relates generally to computer systems and methods which use parallel processing techniques and, more particularly, to a system and methods for converting any parallel random access machine (PRAM) program into an oblivious parallel random access machine (OPRAM) program that simultaneously (i.e, in parallel) processes data while maintaining secure
20 communication in the presence of third party adversaries.

BACKGROUND OF THE INVENTION

 Cryptography is directed to secure communication techniques in the presence of third parties, known as adversaries. More generally, cryptography
25 includes constructing and analyzing protocols that block adversaries to ensure data confidentiality, data integrity, authentication, and non-repudiation.

 Completeness results in cryptography provide general transformations from arbitrary functionalities described in a particular computational model, to solutions for executing the functionality securely within a desired adversarial
30 model. Certain previous results modeled computation as Boolean circuits, and showed how to emulate the circuit securely gate by gate.

 As the complexity of modern computing tasks scales at tremendous rates, it has become clear that the circuit model is not appropriate. In particular, converting "lightweight" optimized programs first into a circuit in order to obtain
35 security is not a viable option. Large effort has recently been focused on enabling

direct support of functionalities modeled as Turing machines or random-access machines (RAM). This approach avoids several sources of expensive overhead in converting modern programs into circuit representations. However, it actually introduces a different dimension of inefficiency. RAM (and single-tape Turing)

5 machines do not support parallelism. Thus, even if an insecure program can be heavily parallelized, its secure version will be inherently sequential.

Modern computing architectures are better captured by the notion of a Parallel RAM (PRAM). In the PRAM model of computation, several (polynomially many) CPUs are simultaneously running, accessing the same shared "external"

10 memory. It should be noted that PRAM Central Processing Units (CPUs) can model physical processors within a single multicore system, as well as distinct computing entities within a distributed computing environment.

A machine is said to be memory oblivious, or simply oblivious, if the sequences of memory accesses made by the machine on two inputs with the

15 same running time are identically (or close to identically) distributed. It has been previously shown that a Turing machine can be compiled into an oblivious one with only a logarithmic slowdown in running-time. Roughly ten years later, the notion of Oblivious RAM (ORAM) was proposed, and showed a similar transformation result with polylogarithmic slowdown. In recent years, ORAM

20 compilers have become a central tool in developing cryptography for RAM programs, and a great deal of research has gone toward improving both the asymptotic and concrete efficiency of ORAM compilers. However, for all such compilers, the resulting program is inherently sequential.

ORAM lies at the base of a wide range of cryptographic applications such

25 that parallelism within the corresponding secure application is desired. Hiding correlated lookups while maintaining efficiency is perhaps the core challenge in building oblivious RAMs. In order to bypass this problem, ORAM compilers may heavily depend on the ability of the CPU to move data around, and to update its secret state after each memory access. However, in the parallel setting, having

30 all processors attempt to perform a lookup directly within a standard ORAM construction corresponds to running the ORAM several times without moving data or updating state, which immediately breaks security in all existing ORAM compiler constructions. Furthermore, most cannot afford for the CPUs to take turns accessing and updating the data sequentially.

Therefore, there is a need to formulate cryptographic primitives that directly support PRAM computations while ensuring that secret information is not leaked via the memory access patterns of the resulting program execution.

5 SUMMARY OF THE INVENTION

According to the invention, any PRAM is converted into an oblivious PRAM (OPRAM), while only inducing a polylogarithmic slowdown to both the total and parallel complexities of the program. More specifically, the invention comprises a compiler that takes any PRAM and converts it into one whose
 10 distribution of memory accesses is statistically independent of the data (with negligible error).

The invention is built on a sequential ORAM compiler with a binary tree-based structure. At a high level, data is stored in the structure of a binary tree, where each node of the tree corresponds to a fixed-size bucket that may hold a
 15 collection of data items. Each memory cell **addr** in the original database is associated with a random path (equivalently, leaf) within a binary tree, as specified by a position map $\text{path}_{\text{addr}} = \text{Pos}(\text{addr})$.

The schemes maintain three invariants: (1) the content of memory cell **addr** is found in one of the buckets along the path $\text{path}_{\text{addr}}$, (2) given the view of
 20 the adversary (i.e., memory accesses) up to any point in time, the current mapping Pos appears uniformly random, and (3) with overwhelming probability, no node in the binary tree will ever “overflow” in the sense that its corresponding memory bucket is instructed to store more items than its fixed capacity.

These invariants are maintained by the following general steps: (1)
 25 Lookup: To access a memory item **addr**, the CPU accesses all buckets down the path $\text{path}_{\text{addr}}$, and removes it where found, (2) Data “put-back”: At the conclusion of the access, the memory item **addr** is assigned a freshly random path $\text{Pos}(\text{addr}) \leftarrow \text{path}'_{\text{addr}}$, and is returned to the root node of the tree, and (3) Data flush: To ensure the root (and any other bucket) does not overflow, data is
 30 “flushed” down the tree, for example, by selecting and emptying two random buckets from each level into their appropriate children, or by choosing an independent path in the tree and pushing data items down this path as far as they will go.

In tree-based ORAMs if CPUs access different data items in a time step, they access different paths in the tree, whereas if they attempt to simultaneously access the same data item, they will each access the same path in the tree, blatantly revealing this collision. The invention solves this problem by inserting a CPU-coordination phase that lets the CPUs check-through an oblivious aggregation operation whether two (or more) of them wish to access the same data item. If so, a representative is selected, for example the CPU with the smallest id, to actually perform the memory access, and all the others merely perform “dummy” lookups. Finally, the representative CPU communicates the read value back to all the other CPUs that wanted to access the same data item using an oblivious multi-cast operation. The challenge is in doing so without introducing too much overhead – namely, allowing only (per-CPU) memory, computation, and parallel time polylogarithmic in both the database size and the number of CPUs – and that itself retains memory obliviousness.

For parallel “put-backs”, after a memory cell is accessed, the (possibly updated) data is assigned a fresh random path and is reinserted to the tree structure. To maintain the required invariants listed above, the item must be inserted somewhere along its new path, without revealing any information about the path. In tree-based ORAMs, this is done by reinserting at the root node of the tree. However, this single node can hold only a small bounded number of elements (corresponding to the fixed bucket size), whereas the number of processors m – each with an item to reinsert – may be significantly larger. To overcome this problem, instead of returning data items to the root, they are directly inserted into level $\log m$ of the tree, while ensuring that they are placed into the correct bucket along their assigned path. Note that level $\log m$ contains m buckets, and since the m items are each assigned to random leaves, each bucket will in expectation be assigned exactly 1 item. The challenge in this step is specifying how the m CPUs can insert elements into the tree while maintaining memory obliviousness. For example, if each CPU simply inserts its own item into its assigned node, information is immediately leaked about its destination leaf node. To resolve this issue, the CPUs obviously route items between each other, so that eventually the i th CPU holds the items to be insert to the i th node,

and all CPUs finally perform either a real or a dummy write to their corresponding node.

To prevent overflows and ensure that no new overflows are introduced after inserting m items, m flushes are performed instead of once, and all these m flushes are done in parallel: each CPU simply performs an independent flush. These parallel flushes may lead to conflicts in nodes accessed (e.g., each flush operation likely accesses the root node). As before, this issue is resolved by having the CPUs elect some representative to perform the appropriate operations for each accessed node: however, that this step is required only for correctness, and not for security.

With ORAM lying at the base of a wide range of cryptographic applications, replacing the underlying ORAM with an OPRAM immediately enables parallelism within the corresponding secure application.

As one example, garbled circuits allow a user to convert a circuit and an input into garbled versions in such a way that the garbled circuit can be evaluated on garbled input to reveal an output, but without revealing further information on the circuit or the input. Garbling schemes have found countless applications in cryptography, ranging from delegation of computation to secure multi-party protocols. According to the invention, any OPRAM compiler can be used to attain garbled PRAM, where the time to generate and evaluate the garbled PRAM program scales with the parallel time complexity of the program. In this embodiment, the compiler construction is an extension of the scheme based on identity-based encryption (IBE). According to the invention, the garbled PRAM construction yields constant-round secure protocols where the time to execute the protocol scales with the parallel time of the program being evaluated

Another example of the invention is directed to improved, parallelized outsourced data. Standard ORAM has been shown to yield effective, practical solutions for securely outsourcing data storage to an untrusted server. Efficient OPRAM compilers enable these systems to support secure efficient parallel accesses to outsourced data. Specifically, OPRAM procedures securely aggregate parallel data requests and resolve conflicts client-side, minimizing expensive client-server communications. As network latency is a major bottleneck in ORAM implementations, such parallelization may yield significant improvements in efficiency.

In a similar example, use of OPRAM further enables secure access and manipulation of outsourced shared data by multiple (mutually trusting) clients, i.e., multi-client outsourced data. Each client can simply act as an independent CPU, and executes the OPRAM-compiled program corresponding to the parallel concatenation of their independent tasks.

As another example, relying instead on OPRAM opens the door to achieving secure hardware in the multi-processor setting. According to the invention, secure multi-processor architectures are implemented by using OPRAM to prevent information leakage via access patterns of the secure processor to the potentially insecure memory.

Yet another application of the invention is directed to secure two-party and multi-party computation of PRAMs. Secure multi-party computation (MPC) enables mutually distrusting parties to jointly evaluate functions on their secret inputs without revealing information on the inputs beyond the desired function output. OPRAM has become a central tool in achieving efficient MPC protocols for securely evaluating RAM programs. By instead relying on OPRAM, these protocols can leverage parallelizability of the evaluated programs. In one particular embodiment, OPRAM can reduce the round complexity of existing two-party computation protocols for RAMs to the parallel (and not sequential) complexity of the evaluated program. An advantage of the invention is that large data sets can be securely and privately processed over multiple parties with parallel, distributed procedures as described more fully in the publication entitled "*Large-Scale Secure Computation: Multi-party Computation for (Parallel) RAM Programs*" dated June 6, 2015 authored by Boyle et al., incorporated by reference.

The invention is directed to an expressive model where the number of active CPUs may vary over time (as long as the pattern of activation is fixed a priori). In this sense, PRAMs efficiently capture the best of both RAM and the circuit models (PRAM processor activation patterns can be aligned to match varying width circuit topology).

The invention and its attributes and advantages may be further understood and appreciated with reference to the detailed description below of one contemplated embodiment, taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

The preferred embodiments of the invention will be described in conjunction with the appended drawings provided to illustrate and not to limit the invention, where like designations denote like elements, and in which:

FIG. 1 illustrates an exemplary network system that may be used to implement the invention.

FIG. 2 illustrates an exemplary computer system that may be used to implement the invention.

FIG. 3 illustrates a block diagram of certain functions according to the invention.

FIG. 4 illustrates a block diagram of a software compiler according to the invention.

FIG. 5 illustrates pseudocode for an oblivious parallel data access procedure according to the invention.

FIG. 6 illustrates pseudocode for combining processor instructions to perform an update bucket procedure according to the invention.

FIG. 7 illustrates pseudocode for combining processor instructions to perform an update bucket procedure according to another embodiment of the invention.

FIG. 8 illustrates a graph of the procedure illustrated in FIG. 6 according to the invention.

FIG. 9 is a flowchart of steps of the oblivious aggregation operation according to the invention.

FIG. 10 illustrates pseudocode for a space-efficient oblivious data aggregation procedure according to the invention.

FIG. 11 is a flowchart of certain steps of the oblivious aggregation operation according to the invention.

FIG. 12 is a flowchart of steps of the oblivious data multi-cast operation according to the invention.

FIG. 13 illustrates pseudocode for a space-efficient oblivious data multi-casting procedure according to the invention.

FIG. 14 is a flowchart of steps of the oblivious data route operation according to the invention.

FIG. 15 illustrates pseudocode for a space-efficient oblivious data routing procedure according to the invention.

FIG. 16 illustrates trimming of the OPRAM construction according to the invention.

5

DETAILED DESCRIPTION OF EMBODIMENTS OF THE INVENTION

FIG. 1 illustrates an exemplary communications network system 100 that may be used to implement the invention, for example as computer code processed by a processor such as one similar to that described in reference to
10 FIG. 2.

Specifically, the communications network system 100 includes at least one client computer 110A (also referred to as “client”). The client computer 110A may be any device through the use of which a distributed computing environment may be accessed to perform the invention as detailed below, for example, a traditional
15 computer, portable computer, handheld device, mobile phone, personal digital assistant, smart hand-held computing device, cellular telephone, or a laptop or netbook computer, hand held console or MP3 player, tablet, or similar hand held computer device, such as an iPad®, iPad Touch® or iPhone®. More specifically, the client computer 110A may include one or more components as described in
20 reference to the computer system of FIG. 2.

The one or more client computers 110A establish communication with the Internet 120 (including untrusted web browsers) to one or more servers 130 (also referred to as “server”). A server computer 130 permits access to a collection of computing resources and components that can be invoked to instantiate a
25 machine, process, or other resource for a limited or defined duration. For example, one group of resource servers can host and serve an operating system or components thereof to deliver and instantiate a virtual machine. Another group of resource servers can accept requests to host computing cycles or processor time, to supply a defined level of processing power for a machine or virtual
30 machine.

One embodiment of the client computer 110A or server computer 130 may be shown by the exemplary computer system 200 of FIG. 2. Computer system 200 includes an input/output display interface 202 connected to communication infrastructure 204 – such as a bus – that forwards data including graphics, text,

and information, from the communication infrastructure 204 to other components of the computer system 200. The input/output display interface 202 may be, for example, a display device, a keyboard, touch screen, joystick, trackball, mouse, monitor, speaker, printer, Google Glass[®] unit, web camera, any other computer peripheral device, or any combination thereof, capable of entering and/or viewing data.

Computer system 200 includes one or more processors 206 (also referred to as “CPU”), which may be a special purpose or a general-purpose digital signal processor configured to process certain information. Computer system 200 also includes non-transitory computer-readable storage medium such as a main memory 208, for example random access memory, read-only memory, mass storage device, or any combination thereof. Computer system 200 may also include a secondary memory 210 such as a hard disk unit 212, a removable storage unit 214, or any combination thereof. Computer system 200 may also include a communication interface 216, for example, a modem, a network interface (such as an Ethernet card or Ethernet cable), a communication port, a PCMCIA slot and card, wired or wireless systems (such as Wi-Fi, Bluetooth, Infrared), local area networks, wide area networks, intranets, etc.

It is contemplated that the main memory 208, secondary memory 210, communication interface 216, or a combination thereof, function as a non-transitory computer-readable storage medium to store and/or access computer software including computer instructions. Certain embodiments of a computer readable storage medium do not include any transitory signals or waves. For example, computer programs or other instructions may be loaded into the computer system 200 such as through a removable storage device, for example, ZIP disks, magnetic tape, portable flash drive, optical disk such as a CD or DVD or Blu-ray, Micro-Electro-Mechanical Systems (MEMS), nanotechnological apparatus. Specifically, computer software including computer instructions may be transferred from the removable storage unit 214 or hard disc unit 212 to the secondary memory 210 or through the communication infrastructure 204 to the main memory 208 of the computer system 200.

Communication interface 216 allows software, instructions and data to be transferred between the computer system 200 and external devices or external networks. Software, instructions, and/or data transferred by the communication

interface 216 are typically in the form of signals that may be electronic, electromagnetic, optical or other signals capable of being sent and received by the communication interface 216. Signals may be sent and received using wire or cable, fiber optics, a phone line, a cellular phone link, a Radio Frequency (RF)

5 link, wireless link, or other communication channels.

Computer programs, when executed, enable the computer system 200, particularly the processor 206, to implement the invention according to computer software including instructions.

10 The computer system 200 of FIG. 2 is provided only for purposes of illustration, such that the invention is not limited to this specific embodiment. It is appreciated that a person skilled in the relevant art knows how to program and implement the invention using any network architecture or computer system including that known as a cloud computer system.

15 The invention is also directed to computer products, otherwise referred to as computer program products. Computer products store software on any computer useable medium, known now or in the future. Such software, when executed, may implement the methods according to certain embodiments of the invention.

20 FIG. 3 illustrates a block diagram of certain functions 225 according to an embodiment of the invention. As shown in FIG. 3, a source program 250 is stored, for example, in memory 210 such as the hard disk unit 212 (see FIG. 2). A compiler 300 may be stored in the hard disk unit 212 and loaded into the main memory 208 by a function of the operating system. The compiler 300, which performs compilation processing according to the invention, is stored in the hard
25 disk unit 212 and loaded into the main memory 208 by a function of the operating system to convert the source program 202 to executable code in the form of an object program 350. Particularly, the compiler 300 is composed of an oblivious aggregation operation 312, oblivious multi-cast operation 314, and oblivious route operation 316 as subsequently described in reference to FIG. 4. The executable
30 code generated by the compiler 300 is preferably stored in the hard disk drive memory 210 and executed by the operating system.

FIG. 4 illustrates a block diagram of a software compiler 300 composed of a oblivious aggregation operation 312, oblivious multi-cast operation 314, and oblivious route operation 316, all performed in an oblivious fashion – the

sequences of memory accesses made by the machine on two inputs with the same running time are identically (or close to identically) distributed. In other words, the access pattern of a memory-restricted client is hidden on a server with larger memory. The oblivious aggregation operation 312 allows CPUs to

5 efficiently identify a single representative processor and aggregate relevant instructions as described further in reference to FIG. 9, FIG. 10, FIG. 11.

The oblivious multi-cast operation 314 permits the representative processor to communicate information or data such as the the read value of the data item to all relevant requesting processors as described further in reference

10 to FIG. 12, FIG. 13.

The oblivious route operation 316 permits each CPU to insert its currently held data item into the tree structure that lies along a freshly sampled random path, without revealing information on the path as described more fully in reference to FIG. 14, FIG. 15, FIG. 16. For example, routing a data item between

15 the two or more processors occurs until an i th processor holds the data item to be inserted to an i th node such that the data item is assigned a new position in the binary tree structure to obtain a new node comprising a corresponding path and the data item is inserted at the new node or along the corresponding path without revealing any information about the new node and the corresponding

20 path to obtain a write value of the data item.

A solution for a simplified cast is first presented, where there is no concern for minimizing communication between CPUs or the size of required CPU local memory. In such setting, communicating and aggregating information between all CPUs is “for free”.

FIG. 5 illustrates pseudocode for an oblivious parallel data access procedure (with no concern for per-round bandwidth/memory). In particular, each **Access(r,v)** operation is replaced by the modified procedure as defined in FIG. 5. (Here, “broadcast” means to send the specified message to all other processors). An **UpdateBuckets** operation is performed in Step 4 directed to

25 removing old data from the ORAM database and step 6 directed to flushing the ORAM database.

30

FIG. 6 illustrates pseudocode for combining processor instructions to perform an update bucket procedure according to the invention. As shown in

FIG. 6, the **UpdateBuckets** operation combines CPUs' instructions for buckets for implementation by a single representative CPU. This is used for correctness, not security. A formal description of a full-fledged update bucket procedure can be found in FIG. 7.

5 FIG. 8 illustrates a graph of the procedure illustrated in FIG. 6 according to the invention. CPUs 1-3 each wish to modify nodes along their paths as seen in FIG. 8. For each overlapping node, the CPU with lowest id receives and implements the aggregated commands for the node.

FIG. 9 is a flowchart of steps 400 of the oblivious aggregation operation.
10 In particular, when a simultaneous request is made by more than one processor to access a data item at step 402, instructions of all processors are aggregated at step 404 and a processor is selected at step 405 to be a representative processor. At step 406, the representative processor accesses a memory to obtain a read value of the data item stored in a node of a binary tree structure. It
15 is contemplated that each node retains a plurality of data items. At this same time, the other processors perform a dummy read operation at step 407 of accessing the memory to obtain the read value of the data item. The representative or selected processor records the read value of the data item at step 408. Defining a representative or selected processor may also be done to
20 perform an operation within the binary tree structure when a conflict occurs.

FIG. 10 illustrates pseudocode for a space-efficient oblivious data aggregation procedure according to the invention. To achieve oblivious data aggregation, the CPUs sort their data list with respect to the corresponding key values as shown by step 410 of FIG. 11, which illustrates a flow chart of a portion
25 of the oblivious aggregation procedure according to the invention. Each CPU i holds a set of distinct keys such that any data or subset of data associated with the same key can be aggregated by an aggregation function **Agg**. This can be achieved via an implementation of a $\log(m)$ -depth sorting network, and provides the useful guarantee that all data pertaining to the same key are necessarily held
30 by an block of adjacent CPUs. As shown by step 412 of FIG. 11, data is passed among CPUs in a sequence of $\log(m)$ steps such that at the conclusion the "left-most" (i.e., lowest indexed) CPU in each key-block learns the aggregation of all data pertaining to this key. Explicitly, in each step i , each CPU sends all held

information to the CPU 2^i to the “left” of itself, and simultaneously accepts any received information pertaining to its key. At step 414, each CPU learns whether it is the “left-most” representative in each key-block, by simply checking whether its left-hand neighbor holds the same key. From here, the CPUs have succeeded
 5 in aggregating information for each key at a single representative CPU. Lastly, the original sorting procedure is reversed to return the aggregated information to one of the CPUs who originally requested it, i.e., the oblivious data multi-cast operation according to the invention described more fully below.

FIG. 12 is a flowchart of steps of the oblivious data multi-cast operation
 10 according to the invention. In particular, the representative processor communicates information or data such as the the read value of the data item to all relevant requesting processors. Namely, a subset of processors must deliver information to (unknown) collections of other processors who request it. As can be seen in FIG. 12, the original sorting procedure described in reference to FIG.
 15 9, FIG. 10 and FIG. 11 is reversed to multi-cast the aggregated information to one of the CPUs who originally requested it. At step 452 the CPUs sort their data list with respect to the corresponding key values. Data is passed among CPUs in a sequence of $\log(m)$ steps such that at the conclusion the “right-most” (i.e., highest indexed) CPU in each key-block learns the aggregation of all data
 20 pertaining to this key as shown by step 454. Explicitly, in each step i , each CPU sends all held information to the CPU 2^i to the “right” of itself, and simultaneously accepts any received information pertaining to its key. Each CPU learns whether it is the “right-most” representative in each key-block, by simply checking whether its right-hand neighbor holds the same key in order to communicate or multi-cast
 25 information for each key to all requesting CPUs at step 456. FIG. 13 illustrates pseudocode for a space-efficient oblivious data multi-casting procedure according to the invention.

FIG. 14 is a flowchart of steps 470 of the oblivious route operation that enables each CPU to insert its currently held data item into the bucket at level $\log$
 30 m of the tree that lies along a freshly sampled random path, without revealing information on the path. In particular, data items are delivered to their target locations via a fixed-topology routing network. Namely, as shown by step 472, the m processors $CPU_1, \dots, CPU_m$, first writes the relevant m data items msg_i (and

their corresponding destination addresses addr_i) to memory in fixed order, and then rearranges them at step 474 in $\log m$ sequential rounds in order to locate them at the proper locations via the routing network as shown by step 476. At the conclusion of the routing procedure, each node j holds all messages msg_i for which $\text{addr}_i = j$.

For simplicity, it is assumed $m = 2^\ell$ for some $\ell \in \mathbb{N}$. The routing network has depth ℓ in each level $t = 1, \dots, \ell$, each node communicates with the corresponding node whose id agrees in all bit locations except for the t -th (corresponding to its t -th neighbor in the $\log m$ -dimensional Boolean hypercube). These nodes exchange messages according to the t -th bit of their destination addresses addr_i , formally described in FIG. 15. After the t -th round, each message msg_i is held by a party whose id agrees with the destination address addr_i in the first t bits. Therefore, at the conclusion of ℓ rounds, all messages are properly delivered. Again, FIG. 15 illustrates pseudocode for an oblivious data routing procedure according to the invention in which a fixed-topology routing network delivers m messages originally held by m processors to their corresponding destination address within $[m]$.

The “Heavy-OPAccess” structure of the OPRAM presented above in reference to FIG. 5 within the simplified “free CPU communication” setting is combined together with the (oblivious) Route, OblivAgg, and OblivMCast procedures constructed above. In one embodiment of the invention, the number of CPUs m is fixed; however, it can be modified in a straightforward fashion to the more general case (as long as the activation schedule of CPUs is a-priori fixed and public). In order to ensure the Position Map information (associating data items to paths in the tree, and which itself, must be accessed obliviously) is smaller than the database itself, data items are divided into contiguous blocks of constant size α ; blocks are each stored as one unit, are assigned a single path in the tree. This yields a Position Map of size n/α , whose access patterns can then be protected via a recursive application of the original ORAM scheme.

In one embodiment of an OPRAM compiler O , each $\text{Access}(\mathbf{r}, \mathbf{v})$ operation is replaced by a sequence of operations defined by subroutine $\text{OPAccess}(\mathbf{r}, \mathbf{v})$,

which is constructed over the following subsections. The **OPAccess** procedure begins with m CPUs, each with a requested data cell r_i (within some α -block b_i) and some action to be taken (either $\perp$ to denote read, or v_i to denote rewriting cell r_i with value v_i).

- 5 For conflict resolution, **OblivAgg** is run on inputs to select a unique representative $\text{rep}(b_i)$ for each queried block b_i and aggregate all CPU instructions for this b_i .

For recursive access to that position map, each representative CPU samples a fresh random in the tree and performs a (recursive) Read/Write
 10 access command on the position map database to fetch the current position map value ℓ for block b_i and rewrite it with the newly sampled value. Each dummy CPU performs an arbitrary dummy access.

For look up of current memory values, each CPU fetches memory from the database nodes down the path to leaf. When b_i is found, it copies its value v_i
 15 into local memory. Each dummy CPU chooses a random path and makes analogous dummy data fetches along it, ignoring all read values (again, simultaneous data reads do not yield conflicts).

Old data is then removed. For each level in the tree, instructions are aggregated across CPUs accessing the same “buckets” of memory
 20 (corresponding to nodes of the tree) on the server side. Each representative CPU begins with the instruction of “remove block b if it occurs” and dummy CPUs hold the empty instruction (aggregation is as before, but at bucket level instead of the block level). For each bucket to be modified, the CPU with the smallest id from those who wish to modify it executes the aggregated block-removal instructions
 25 for the bucket (again, this aggregation step is purely for correctness and not security).

The updated data is then inserted into the database in parallel and the ORAM database is flushed. In parallel, each CPU initiates an independent flush
 30 of the ORAM tree – this corresponds to selecting a random path down the tree, and pushing all data blocks in this path as far as they will go. To implement the simultaneous flush commands, as before, commands are aggregated across CPUs for each bucket to be modified, and the CPU with the smallest id performs

the corresponding aggregated set of commands. For example, all CPUs wish to access the root node in their flush; the aggregation of all corresponding commands to the root node data is executed by the lowest-numbered CPU who wishes to access this bucket.

- 5 Output is returned by running **OblivMCast** on inputs to communicate the original (pre-updated) value of each data block b_i to the subset of CPUs that originally requested it.

 In the case that the number of CPUs m is fixed and known a priori, the OPRAM construction can be directly trimmed in two places. FIG. 16 illustrates
10 trimming of the OPRAM construction according to the invention. With respect to trimming tops of recursive data trees it is noted that data items are always inserted into the OPRAM trees at level $\log m$, and flushed down from this level. Thus, the top levels in the OPRAM tree are never utilized. In such case, the data buckets in the corresponding tops of the trees, from the root node to level $\log m$
15 for this bound, can simply be removed without affecting the OPRAM. With respect to truncating recursion, in the t -th level of recursion, the corresponding database size shrinks so that oblivious data accesses can be achieved via local CPU communication without needing any tree lookups or further recursion.

 While the disclosure is susceptible to various modifications and alternative
20 forms, specific exemplary embodiments of the invention have been shown by way of example in the drawings and have been described in detail. It should be understood, however, that there is no intent to limit the disclosure to the particular embodiments disclosed, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the scope of the
25 disclosure as defined by the appended claims.

CLAIMS

1. A secure computer program method enabling two or more processors to simultaneously access a shared external memory, the method comprising the steps of:

- 5 selecting a processor from two or more processors to obtain a selected processor and one or more unselected processors, said selecting step occurring when each of the two or more processors require simultaneous access to a data item;
- aggregating by the selected processor instructions of each of the two or
10 more processors to obtain aggregated instructions;
- accessing a memory by the selected processor to obtain a read value of the data item, wherein the data item is stored in a node of a binary tree structure;
- recording by the selected processor the read value of the data item;
- communicating by the selected processor to the one or more unselected
15 processors the read value of the data item;
- routing the data item between the two or more processors until an i th processor holds the data item to be inserted to an i th node such that the data item is assigned a new position in the binary tree structure to obtain a new node comprising a corresponding path; and
- 20 inserting by the selected processor the data item at the new node or along the corresponding path without revealing any information about the new node and the corresponding path to obtain a write value of the data item.

2. The secure computer program method according to claim 1,
25 wherein the selected processor has the smallest id selected from the two or more processors.

3. The secure computer program method according to claim 1,
wherein said accessing step further comprises the step of performing by the one
30 or more unselected processors a dummy read operation of accessing the memory to obtain the read value of the data item.

4. The secure computer program method according to claim 1,
wherein the node retains a plurality of data items.

5. The secure computer program method according to claim 1, wherein the new random path is defined by $\log m$ of the binary tree structure, with m representing a number of processors.

5

6. The secure computer program method according to claim 1, wherein said inserting step further comprises the step of performing by the one or more unselected processors a dummy write operation inserting the data item at the new node or along the corresponding path.

10

7. The secure computer program method according to claim 1 further comprising the step of:

flushing simultaneously by the two or more processors the binary tree structure so that each node of the binary tree structure is not overflowing with data items.

15

8. The secure computer program method according to claim 8 further comprising the step of:

electing a processor from the two or more processors to perform an operation within the binary tree structure when a conflict occurs.

20

9. The secure computer program method according to claim 1, wherein said aggregating step further comprises the steps of:

sorting by a processor a data list according to one or more key values, wherein each processor has an associated data list and each processor holds a block of distinct key values;

25

passing between the two or more processors a sequence of $\log(m)$ steps such that at the conclusion the "left-most" (i.e., lowest indexed) processor in each block aggregates all data pertaining to a particular key to obtain aggregated data; and

30

learning by each processor that it is the "left-most" processor in each block.

10. The secure computer program method according to claim 9, wherein said sorting step further comprises the step of implementing a $\log(m)$ -depth sorting network with all data pertaining to the same key are held by adjacent processors.

5

11. The secure computer program method according to claim 9, wherein said learning step further includes the step of checking by each processor that its left-hand neighbor holds the same key.

10

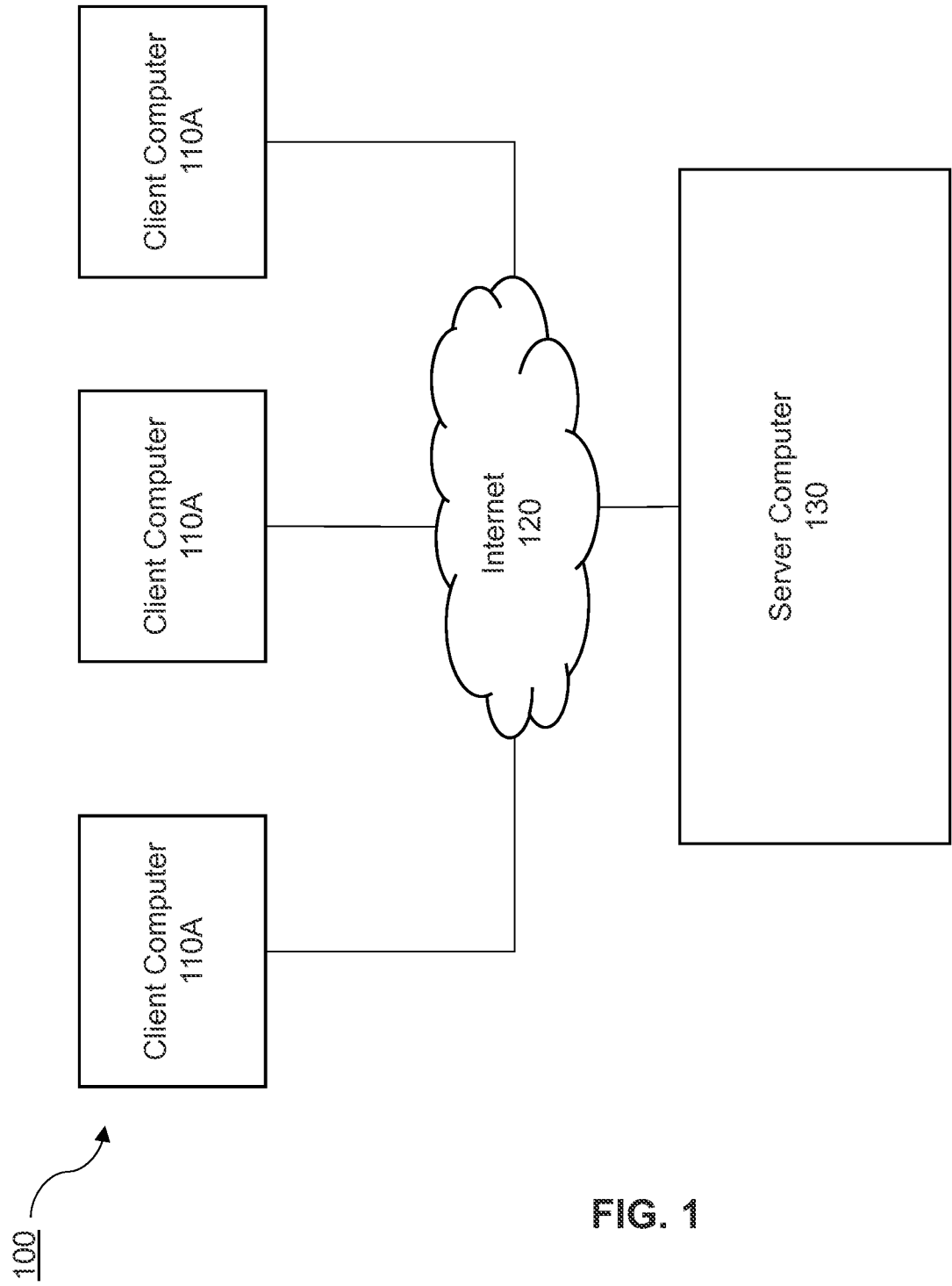


FIG. 1

200 ↗

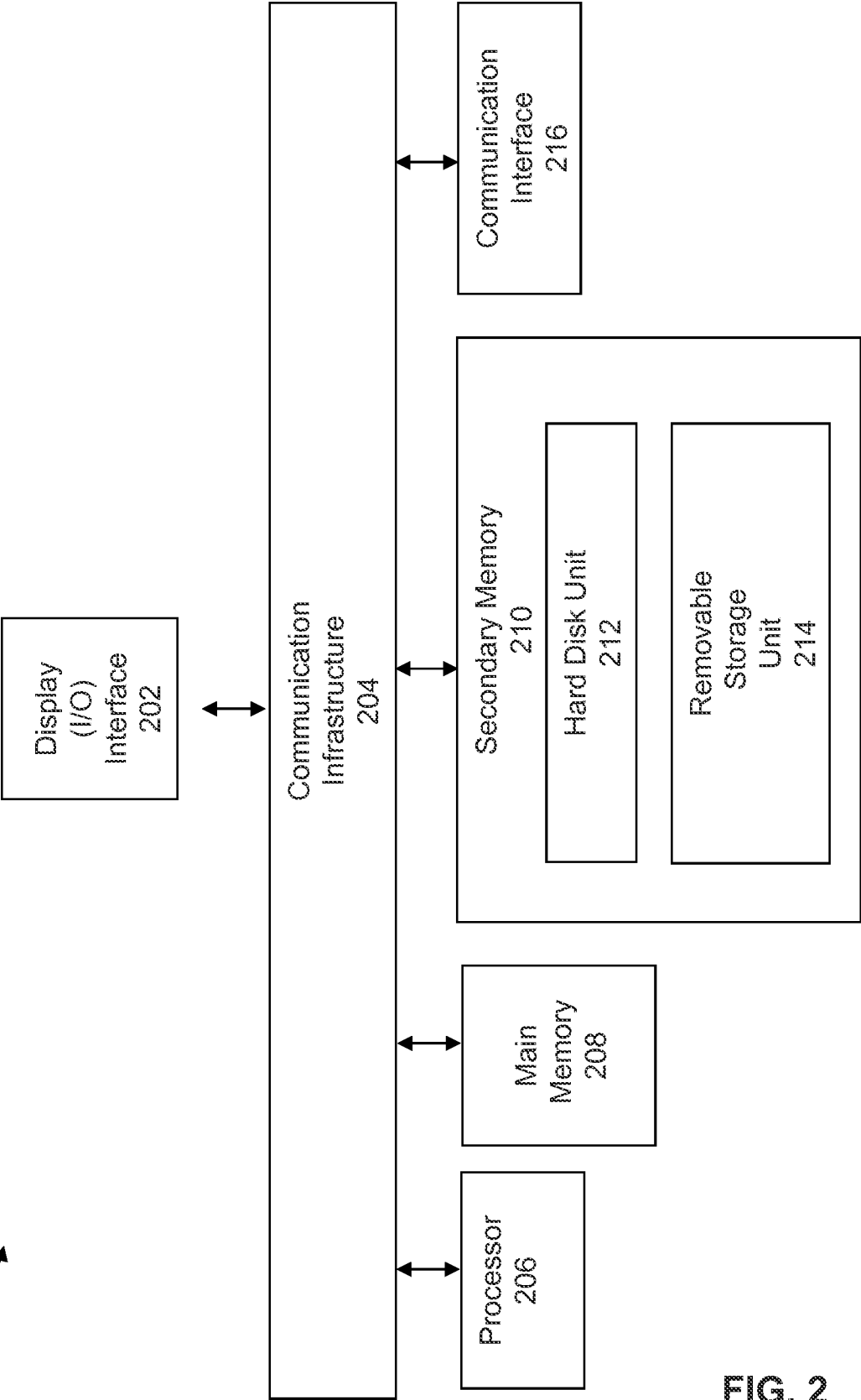
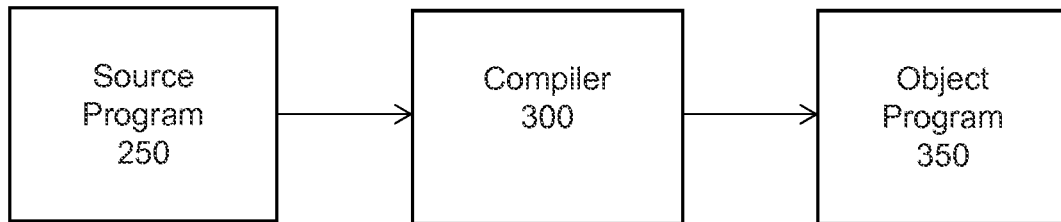
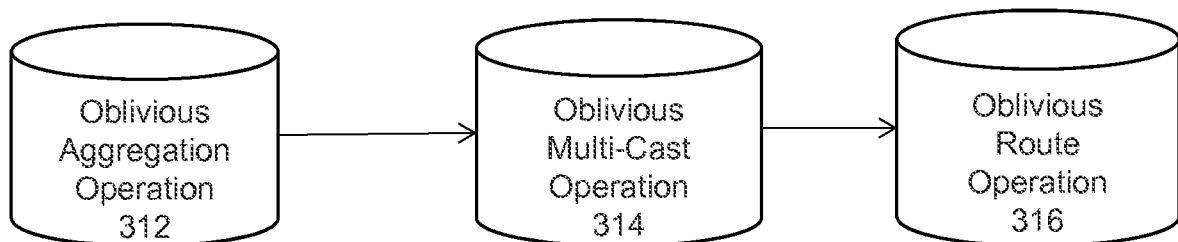


FIG. 2

225**FIG. 3**300**FIG. 4**

Heavy-OPAccess($t, (r_i, v_i)$): The Large Bandwidth Case

To be executed by m processors $CPU_1, \dots, CPU_m$ w.r.t. (recursive) database size $n_t := n/(\alpha^t)$.

Input: Each CPU_i holds: recursion level t , instruction pair (r_i, v_i) with $r_i \in [n_t]$, global parameter α . Each CPU_i performs the following steps, in parallel:

0. Exit Case: If $t \geq \log_\alpha n$, return 0.

This corresponds to requesting the (trivial) position map for a block within a single-leaf tree.

1. Conflict Resolution

(a) Broadcast the instruction pair (r_i, v_i) to all CPUs.

(b) Let $b_i = \lfloor r_i/\alpha \rfloor$. Locally aggregate incoming instructions to block b_i as $\bar{v}_i = \bar{v}_i[1] \cdots \bar{v}_i[\alpha]$, resolving write conflicts (i.e., $\forall s \in [\alpha]$, take $\bar{v}_i[s] \leftarrow v_j$ for minimal j such that $r_j = b_i \alpha + s$). Denote by $\text{rep}(b_i) := \min\{j : \lfloor r_j/\alpha \rfloor = b_i\}$ the smallest index j of *any* CPU whose r_j is in this block b_i . (CPU $\text{rep}(b_i)$ will actually access b_i , while others perform dummy accesses).

2. Recursive Access to Position Map

If $i = \text{rep}(b_i)$: Sample fresh leaf id $\ell'_i \leftarrow [n_t]$. Recurse as $\ell_i \leftarrow \text{Heavy-OPAccess}(t+1, (b_i, \ell'_i))$ to read the current value ℓ_i of $\text{Pos}(b_i)$ and rewrite it with ℓ'_i .

Else: Recursively initiate *dummy* access $x \leftarrow \text{Heavy-OPAccess}(t+1, (1, \perp))$ at arbitrary address (say 1); ignore the read value x . Sample a fresh random leaf id $\ell_i \leftarrow [n_t]$ for a dummy lookup.

FIG. 5A

3. Look Up Current Memory Values

Read the memory contents of all buckets down the path to leaf node ℓ_i defined in the previous step, copying all buckets into local memory.

If $i = \text{rep}(b_i)$: locate and store target block triple (b_i, v_i^{old}, ℓ_i) . Update $\bar{v}$ from Step 1 with existing data: $\forall s \in [\alpha]$, replace any non-written cell values $\bar{v}_i[s] = \emptyset$ with $\bar{v}_i[s] \leftarrow v_i^{old}[s]$. $\bar{v}_i$ now stores the entire data block to be rewritten for block b_i .

4. Remove Old Data from ORAM Database

(a) If $i = \text{rep}(b_i)$: Broadcast (b_i, ℓ_i) to all CPUs. Otherwise: broadcast $(\perp, \ell_i)$.

(b) Initiate UpdateBuckets $(n_i, (\text{remove-}b_i, \ell_i), \{(\text{remove-}b_j, \ell_j)\}_{j \in [m] \setminus \{i\}})$, as in Figure 4.

5. Insert New Data into Database in *Parallel*

(a) If $i = \text{rep}(b_i)$: Broadcast $(b_i, \bar{v}_i, \ell'_i)$, with updated value $\bar{v}_i$ and target leaf ℓ'_i .

(b) Let $\text{lev}^* := \lfloor \log(\min\{m, n_i\}) \rfloor$ be the ORAM tree level with number of buckets equal to number of CPUs (the level where data will be inserted). Locally aggregate all incoming instructions whose path ℓ'_j has lev^* -bit prefix i : $\text{Insert}_i := \{(b_j, \bar{v}_j, \ell'_j) : (\ell'_j)^{(\text{lev}^*)} = i\}$.

(c) Access memory bucket i (at level lev^*) and rewrite contents, inserting data items Insert_i . If bucket i exceeds its capacity, abort with overflow.

FIG. 5B

4. Remove Old Data from ORAM Database
 - (a) If $i = \text{rep}(b_i)$: Broadcast (b_i, ℓ_i) to all CPUs. Otherwise: broadcast $(\perp, \ell_i)$.
 - (b) Initiate UpdateBuckets $(n_i, (\text{remove-}b_i, \ell_i), \{(\text{remove-}b_j, \ell_j)\}_{j \in [m] \setminus \{i\}})$, as in Figure 4.
5. Insert New Data into Database in *Parallel*
 - (a) If $i = \text{rep}(b_i)$: Broadcast $(b_i, \bar{v}_i, \ell'_i)$, with updated value $\bar{v}_i$ and target leaf ℓ'_i .
 - (b) Let $\text{lev}^* := \lfloor \log(\min\{m, n_i\}) \rfloor$ be the ORAM tree level with number of buckets equal to number of CPUs (the level where data will be inserted). Locally aggregate all incoming instructions whose path ℓ'_j has lev^* -bit prefix i : $\text{Insert}_i := \{(b_j, \bar{v}_j, \ell'_j) : (\ell'_j)^{(\text{lev}^*)} = i\}$.
 - (c) Access memory bucket i (at level lev^*) and rewrite contents, inserting data items Insert_i . If bucket i exceeds its capacity, abort with overflow.
6. Flush the ORAM Database
 - (a) Sample a random leaf node $\ell_i^{\text{flush}} \leftarrow [n_i]$ along which to flush. Broadcast ℓ_i^{flush} .
 - (b) If $i \leq n_i$: Initiate UpdateBuckets $(n_i, (\text{flush}, \ell_i^{\text{flush}}), \{(\text{flush}, \ell_j^{\text{flush}})\}_{j \in [m] \setminus \{i\}})$, as in Figure 4.
Recall that flush means to “push” each encountered triple (b, ℓ, v) down to the lowest point at which his chosen flush path and ℓ agree.
7. Update CPUs

If $i = \text{rep}(b_i)$: broadcast the *old* value v_i^{old} of block b_i to all CPUs.

FIG. 5C

UpdateBuckets $(n_t, (\text{mycommand}, \text{mypath}), \{(\text{command}_j, \text{path}_j)\}_{j \in [m] \setminus \{i\}})$

Let $\text{path}^{(0)}, \dots, \text{path}^{(\log n_t)}$ denote the bit prefixes of length 0 (i.e., $\emptyset$) to $\log n_t$ of path.

For each tree level $\text{lev} = 0$ to $\log n_t$, each CPU i does the following at bucket $\text{mypath}^{(\text{lev})}$:

1. Define $\text{CPUs}(\text{mypath}^{(\text{lev})}) := \{i\} \cup \{j : \text{path}_j^{(\text{lev})} = \text{mypath}^{(\text{lev})}\}$ to be the set of CPUs requesting changes to bucket $\text{mypath}^{(\text{lev})}$. Let $\text{bucket-rep}(\text{mypath}^{(\text{lev})})$ denote the *minimal* index in the set.
2. If $i \neq \text{bucket-rep}(\text{mypath}^{(\text{lev})})$, do nothing. Otherwise:

Case 1: $\text{mycommand} = \text{remove-}b_i$.

Interpret each $\text{command}_j = \text{remove-}b_j$ as a target block id b_j to be removed. Access memory bucket $\text{mypath}^{(\text{lev})}$ and rewrite contents, removing any block b_j for which $j \in \text{CPUs}(\text{mypath}^{(\text{lev})})$.

FIG. 6A

Case 2: mycommand = flush.

Define Flush $\subset \{L, R\}$ as $\{v : \exists \text{ path}_j \text{ s.t. } \text{path}_j^{(\text{lev}+1)} = \text{mypath}^{(\text{lev})} \parallel v\}$, associating $L \equiv 0$, $R \equiv 1$. This determines whether data will be flushed left and/or right from this bucket. Access memory bucket $\text{mypath}^{(\text{lev})}$; denote its collection of stored data blocks b by ThisBucket. Partition ThisBucket $\equiv \text{ThisBucket-L} \cup \text{ThisBucket-R}$ into those blocks whose associated leaves continue to the left or right (i.e., $\text{ThisBucket-L} := \{b_j \in \text{ThisBucket} : \bar{c}_j^{(\text{lev}+1)} = \text{mypath}^{(\text{lev})} \parallel 0\}$, and similar for 1).

- If $L \in \text{Flush}$, then set $\text{ThisBucket} \leftarrow \text{ThisBucket} \setminus \text{ThisBucket-L}$, access memory bucket $\text{mypath}^{(\text{lev})} \parallel 0$, and insert data items ThisBucket-L into it.
- If $R \in \text{Flush}$, then set $\text{ThisBucket} \leftarrow \text{ThisBucket} \setminus \text{ThisBucket-R}$, access memory bucket $\text{mypath}^{(\text{lev})} \parallel 1$, and insert data items ThisBucket-R into it.

Rewrite the contents of bucket $\text{mypath}^{(\text{lev})}$ with updated value of ThisBucket. If any bucket exceeds its capacity, abort with overflow.

FIG. 6B

UpdateBuckets(m , (command $_i$, path $_i$))

Let path $^{(1)}$, path $^{(2)}$, ..., path $^{(\log n)}$ denote the bit prefixes of length 1 to $\log n$ of path.

For each level lev = 1, ..., $\log n$ of the tree:

1. The CPUs invoke the oblivious aggregation procedure OblivAgg as follows.

Case 1: command $_i$ = remove- b_i .

Each CPU i sets key $_i$ = path $_i^{(\text{lev})}$ and data $_i$ = b_i if b_i is in the bucket of node $\ell_i^{(\text{lev})}$, and data $_i$ = $\perp$ otherwise. Use the union function $\text{Agg}(\{\text{data}_i\}) = \{b : \exists \text{data}_i = b\}$ as the aggregation function.

Case 2: command $_i$ = flush.

Each CPU i sets key $_i$ = path $_i^{(\text{lev})}$ and data $_i$ = L (resp., R) if the (lev + 1)-st bit of path $_i$ is 0 (resp., 1). Use the union function as the aggregation function.

At the conclusion of the protocol, each bucket path $_i^{(\text{lev})}$ is assigned to a representative CPU bucket-rep(path $_i^{(\text{lev})}$) with aggregated commands agg-command $_i$.

FIG. 7A

2. Each representative CPU performs the updates:

If $i \neq \text{bucket-rep}(\text{path}_i^{(\text{lev})})$, do nothing. Otherwise:

Case 1: $\text{command}_i = \text{remove-}b_i$.

Remove all blocks $b \in \text{agg-command}_i$ in the bucket $\text{path}_i^{(\text{lev})}$ by accessing memory bucket $\text{path}_i^{(\text{lev})}$ and rewriting contents.

Case 2: $\text{command}_i = \text{flush}$.

Access memory buckets $\text{path}_i^{(\text{lev})}$, $\text{path}_i^{(\text{lev})} \parallel 0$, $\text{path}_i^{(\text{lev})} \parallel 1$, perform flush operation locally according to $\text{agg-command}_i \subset \{L, R\}$, and write the contents back.

Specifically, denote the collection of stored data blocks b in $\text{path}_i^{(\text{lev})}$ by ThisBucket . Partition $\text{ThisBucket} = \text{ThisBucket-L} \cup \text{ThisBucket-R}$ into those blocks whose associated leaves continue to the left or right (i.e., $\{b_j \in \text{ThisBucket} : \ell_j^{(\text{lev}+1)} = \text{mypath}^{(\text{lev})} \parallel 0\}$, and similar for 1).

- If $L \in \text{agg-command}_i$, then set $\text{ThisBucket} \leftarrow \text{ThisBucket} \setminus \text{ThisBucket-L}$, and insert data items ThisBucket-L into bucket $\text{path}_i^{(\text{lev})} \parallel 0$.
- If $R \in \text{agg-command}_i$, then set $\text{ThisBucket} \leftarrow \text{ThisBucket} \setminus \text{ThisBucket-R}$, and insert data items ThisBucket-L into bucket $\text{path}_i^{(\text{lev})} \parallel 0$.

FIG. 7B

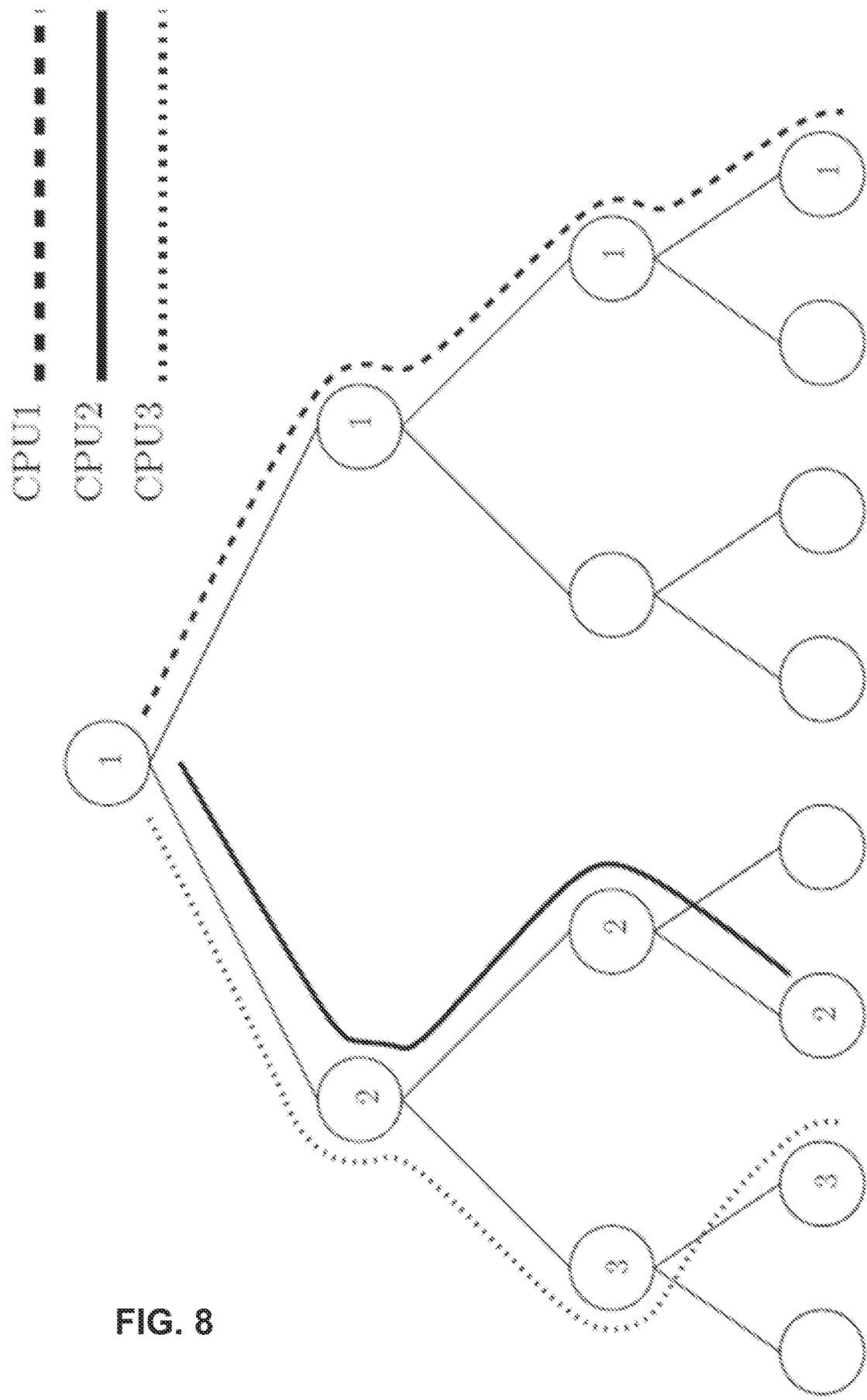


FIG. 8

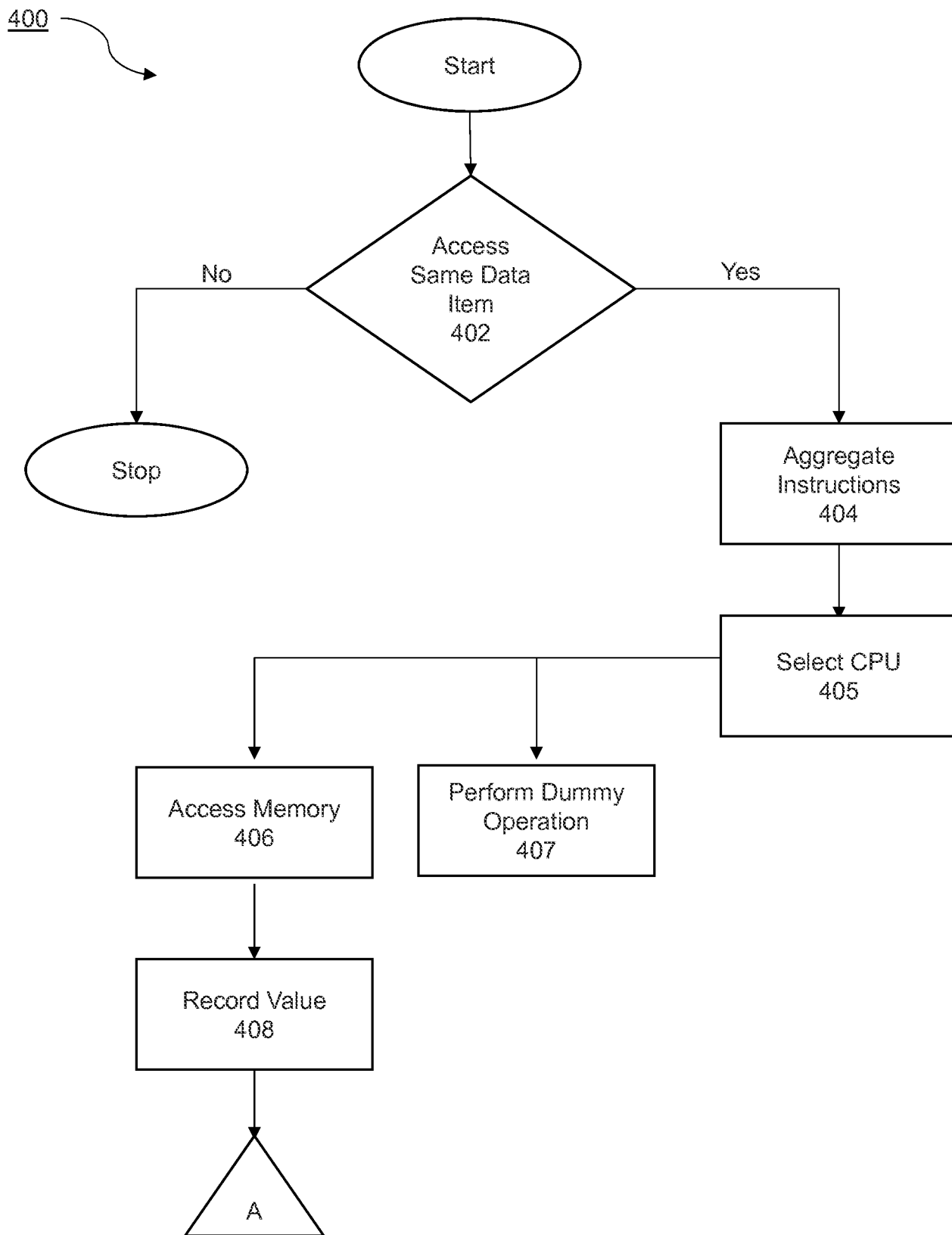


FIG. 9

Oblivious Aggregation Procedure OblivAgg (w.r.t. Agg)

Input: Each CPU $i \in [m]$ holds a pair $(key_i, data_i)$.

Output: Each CPU $i \in [m]$ outputs a triple $(rep_i, key_i, aggdata_i)$ corresponding to either (dummy, $\perp$, $\perp$) or with $aggdata_i = \text{Agg}(\{data_j : key_j = key_i\})$, as further specified in Section 3.1.

1. **Sort on key_i .** Each CPU_i initializes a triple $(sourceid_i, keytemp_i, datatemp_i) \leftarrow (i, key_i, data_i)$.

For each layer $L_1, \dots, L_d$ in the sorting network:

- Let $L_\ell = ((i_1, j_1), \dots, (i_{m/2}, j_{m/2}))$ be the comparators in the current layer ℓ .
- In *parallel*, for each $t \in [m/2]$, the corresponding pair of CPUs (CPU_{i_t}, CPU_{j_t}) perform the following pairwise sort w.r.t. key :

If $keytemp_{j_t} < keytemp_{i_t}$, then

swap $(sourceid_{i_t}, keytemp_{i_t}, datatemp_{i_t}) \leftrightarrow (sourceid_{j_t}, keytemp_{j_t}, datatemp_{j_t})$.

2. **Aggregate to left.** For $t = 0, 1, \dots, \log m$:

- (Pass to left). Each CPU_i for $i > 2^t$ sends his current pair $(keytemp_i, datatemp_i)$ to CPU_{i-2^t} .
- (Aggregate). Each CPU_i for $i < m - 2^t$ receiving a pair $(keytemp_j, datatemp_j)$ will aggregate it into own pair if the keys match. That is, if $keytemp_i = keytemp_j$, then set $datatemp_i \leftarrow \text{Agg}(datatemp_i, datatemp_j)$. In both cases, the received pair is then erased.

The left-most CPU_i with $keytemp_i = key$ now has $\text{Agg}(\{datatemp_j : keytemp_j = key\})$.

FIG. 10A

3. **Identify representatives.** For each value key_j , the left-most CPU i currently holding $\text{keytemp}_i = \text{key}_j$ will identify himself as (temporary) representative.
 - Each CPU_i for $i < m$: send keytemp_i to right-band neighbor, CPU_{i+1} .
 - Each CPU_i for $i > 1$: If the received value keytemp_{i-1} matches his own keytemp_i , then set $\text{rep}_i \leftarrow$ “dummy” and zero out $\text{keytemp}_i \leftarrow \perp$, $\text{datatemp}_i \leftarrow \perp$. Otherwise, set $\text{rep}_i \leftarrow$ “rep”. (CPU_1 always sets $\text{rep}_1 \leftarrow$ “rep”).
4. **Reverse sort (i.e., sort on sourceid_i).** Return aggregated data to a requesting CPU.

For each layer $L_1, \dots, L_d$ in the sorting network:

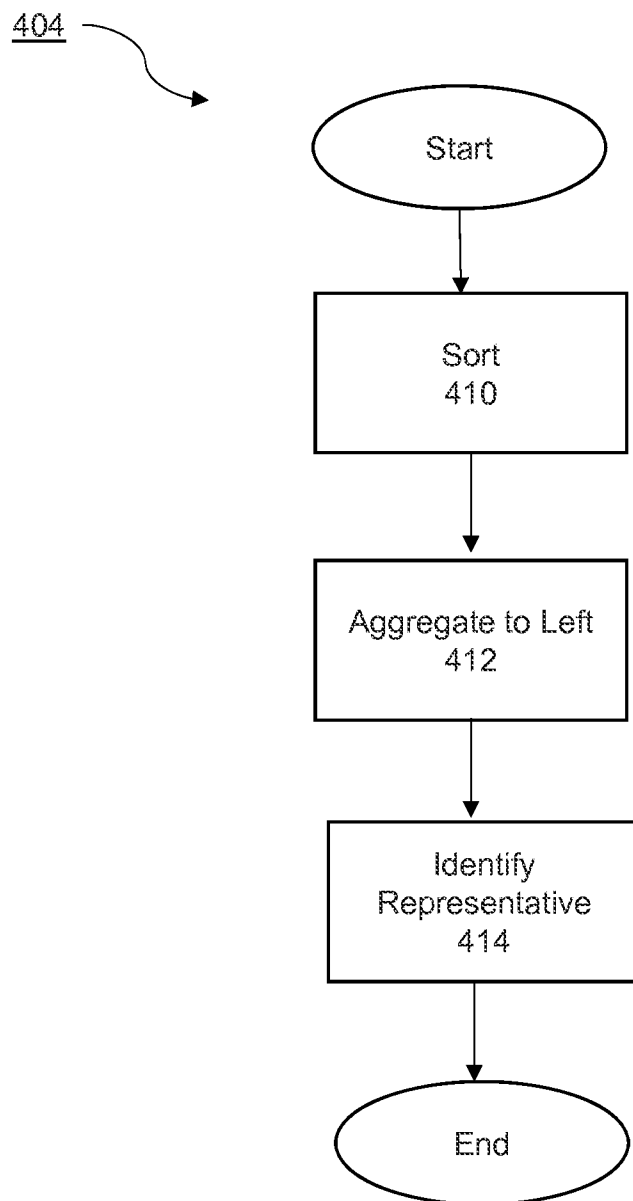
 - Let $L_\ell = ((i_1, j_1), \dots, (i_{m/2}, j_{m/2}))$ be the comparators in the current layer ℓ .
 - Each CPU_i initializes $\text{idtemp} \leftarrow \text{sourceid}_i$. In *parallel*, for each $t \in [m/2]$, the corresponding pair of CPUs ($\text{CPU}_{i_t}, \text{CPU}_{j_t}$) perform the following pairwise sort w.r.t. sourceid :

If $\text{idtemp}_{j_t} < \text{idtemp}_{i_t}$, then

swap ($\text{idtemp}_{i_t}, \text{rep}_{i_t}, \text{keytemp}_{i_t}, \text{datatemp}_{i_t}$) $\leftrightarrow$ ($\text{idtemp}_{j_t}, \text{rep}_{j_t}, \text{keytemp}_{j_t}, \text{datatemp}_{j_t}$).

At the conclusion, each CPU_i holds a tuple ($\text{idtemp}_i, \text{rep}_i, \text{keytemp}_i, \text{datatemp}_i$) with $\text{idtemp}_i = i$ and $\text{keytemp}_i = \text{key}_i$.
5. **Output.** Each CPU_i outputs the triple ($\text{rep}_i, \text{key}_i, \text{datatemp}_i$).

FIG. 10B

**FIG. 11**

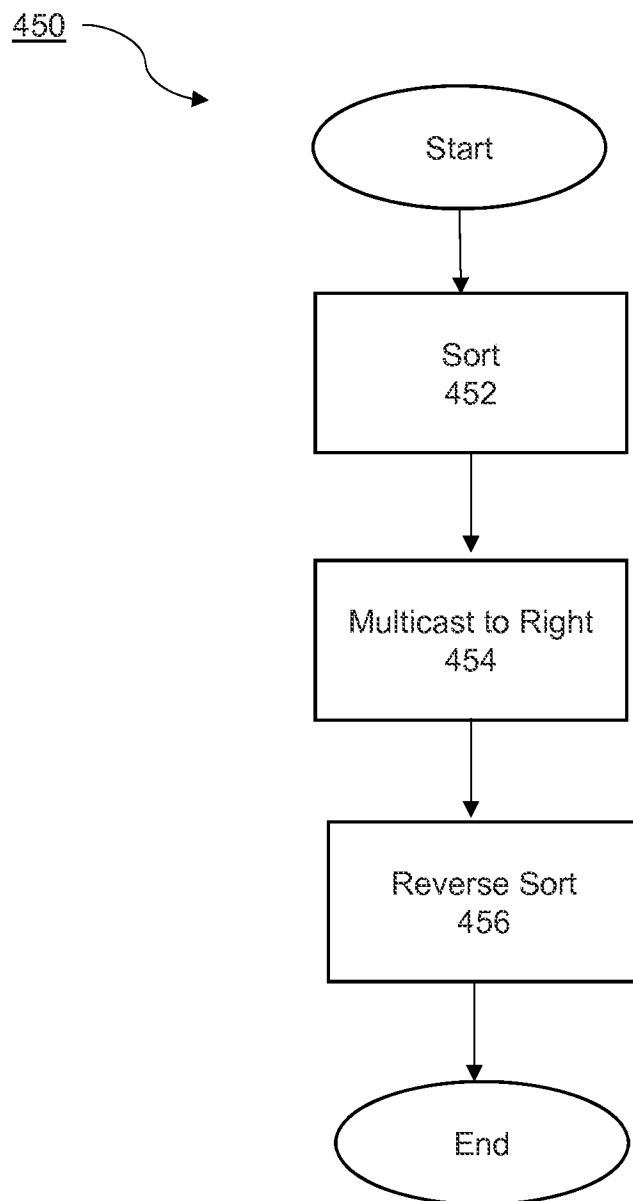


FIG. 12

Oblivious Multicasting Procedure OblivMCast

Input: Each CPU i holds $(key_i, data_i)$ with the following promise. Let $K = \bigcup \{key_i\}$ denote the set of distinct keys. For every $key \in K$, there exists a unique agent i with $key_i = key$ such that $data_i \neq \perp$; let $data_{key}$ denote such $data_i$.

Output: Each agent i outputs $out_i = (key_i, data_{key_i})$.

1. Sort on $(key_i, data_i)$. Each CPU_i initializes $(sourceid_i, keytemp_i, datatemp_i) \leftarrow (i, key_i, data_i)$.

For each layer $L_1, \dots, L_d$ in the sorting network

- Let $L_\ell = ((i_1, j_1), \dots, (i_{m/2}, j_{m/2}))$ be the comparators in the current layer ℓ .
- In *parallel*, for each $t \in [m/2]$, the corresponding pair of CPUs (CPU_{i_t}, CPU_{j_t}) perform the following pairwise sort w.r.t. key , additionally pushing payloads $data_{key}$ to the left:

If (i) $keytemp_{j_t} < keytemp_{i_t}$, or (ii) $keytemp_{j_t} = keytemp_{i_t}$ and $datatemp_{j_t} \neq \perp$, then
 swap $(sourceid_{i_t}, keytemp_{i_t}, datatemp_{i_t}) \leftrightarrow (sourceid_{j_t}, keytemp_{j_t}, datatemp_{j_t})$.

2. Multicast to right. For $t = 0, 1, \dots, \log m$:
 - (Pass to right). Each CPU_i for $i \leq m - 2^t$ sends his current pair $(keytemp_i, datatemp_i)$ to CPU_{i+2^t} .
 - (Aggregate). Each CPU_i for $i > 2^t$ receiving a pair $(keytemp_j, datatemp_j)$ with $j = i - 2^t$ update its data as follows. If $keytemp_i = keytemp_j$ and $datatemp_j \neq \perp$, then set $datatemp_i \leftarrow datatemp_j$.

Every CPU i now holds $(keytemp_i, datatemp_i) = (key, data_{key})$ for some $key \in K$.

FIG. 13A

3. Reverse sort (i.e., sort on sourceid_i). Return received data to an original requesting CPU.

For each layer $L_1, \dots, L_d$ in the sorting network:

- Let $L_\ell = ((i_1, j_1), \dots, (i_{m/2}, j_{m/2}))$ be the comparators in the current layer ℓ .
- Each CPU_i initializes $\text{idtemp} \leftarrow \text{sourceid}_i$. In *parallel*, for each $t \in [m/2]$, the corresponding pair of CPUs (CPU_{i_t}, CPU_{j_t}) perform the following pairwise sort w.r.t. sourceid :

If $\text{idtemp}_{j_t} < \text{idtemp}_{i_t}$, then

swap $(\text{idtemp}_{i_t}, \text{keytemp}_{i_t}, \text{datatemp}_{i_t}) \leftrightarrow (\text{idtemp}_{j_t}, \text{keytemp}_{j_t}, \text{datatemp}_{j_t})$.

At the conclusion, each CPU_i holds a tuple with $(\text{idtemp}_i, \text{keytemp}_i, \text{datatemp}_i)$ with $\text{idtemp}_i \equiv i$, $\text{keytemp}_i \equiv \text{key}_i$, and $\text{datatemp}_i \equiv \text{data}_{\text{key}_i}$.

4. Output. Each CPU_i outputs $\text{output}_i \equiv (\text{key}_i, \text{data}_{\text{key}_i})$.

FIG. 13B

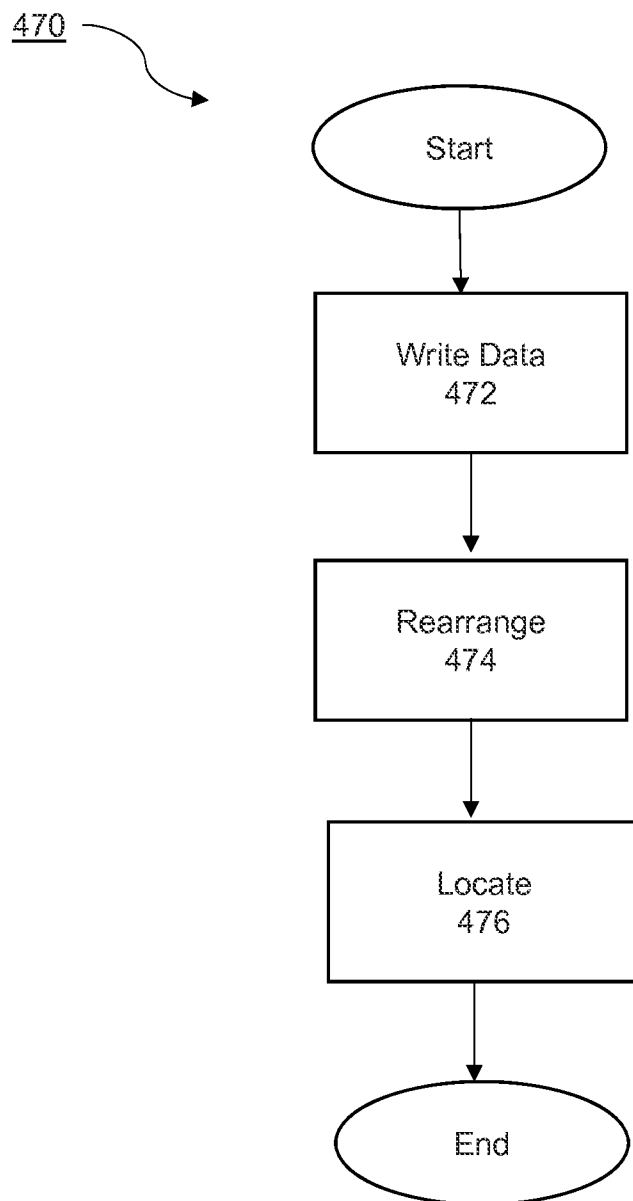
**FIG. 14**

FIG. 15

Parallel Insertion Routing Protocol $\text{Route}(m, (\text{msg}_i, \text{addr}_i))$

Input: CPU_i holds: message msg_i with target destination addr_i , and global threshold K .

Output: CPU_i holds $\{\text{msg}_j : \text{addr}_j = i\}$.

Let $\text{lev}^* = \log m$ (assumed $\in \mathbb{N}$ for simplicity). Each CPU_i performs the following.

Initialize $M_{i,0} \leftarrow \text{msg}_i$. For $t = 1, \dots, \text{lev}^*$:

1. Perform the following symmetric message exchange with $\text{CPU}_{i \oplus 2^t}$:

$$M_{i,t+1} \leftarrow \{\text{msg}_j \in M_{i,t} \cup M_{i \oplus 2^t, t} : (\text{addr}_j)_t = (i)_t\}.$$

2. If $|M_{i,t+1}| > K$ (i.e., memory overflow), then CPU_i aborts.

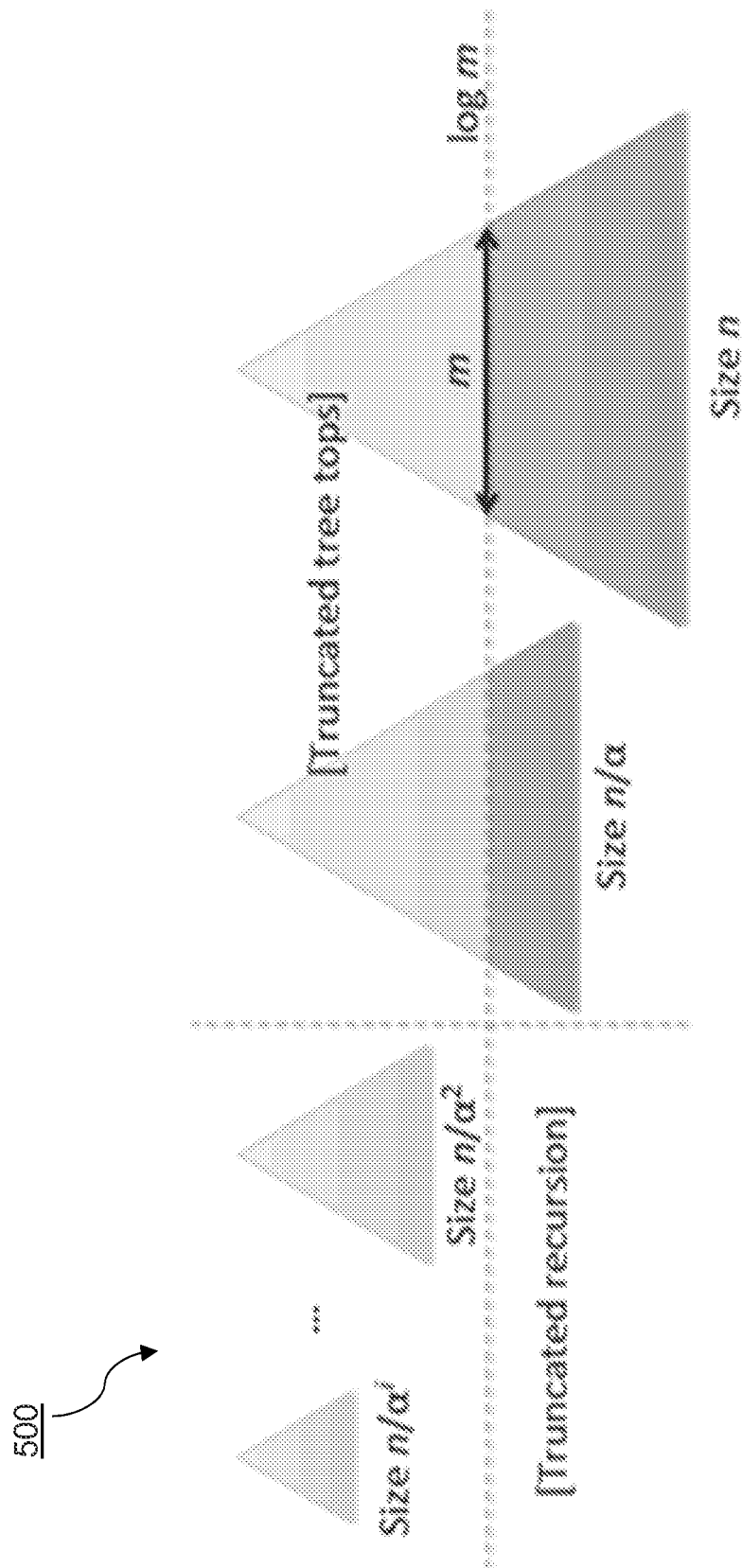


FIG. 16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 15/43210

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 13/00 (2015.01)

CPC - G06F 3/0601

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

CPC: G06F 3/0601; USPC: 711/100; IPC: G06F 13/00 (2015.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC: G06F 3/0601, G06F 3/0659, G11B 27/034, G06F 21/79, G06F 12/0866; USPC: 711/100, 711/104, 711/147; IPC(8): G06F 13/00
(keyword limited - see search terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatBase; Google; Google Scholar

Terms: processor, node, parallel, multiprocessor, memory, storage, oblivious, access, read, write, simultaneous, concurrent, routing, path, tree, address, data, packet, policy, iterate, conflict, threshold, key, sort, order, log, buffer.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2012/0221808 A1 (COON et al.), 30 August 2012 (30.08.2012), entire document, especially Abstract; Para [0008], [0010], [0011], [0033], [0110]	1-11
Y	GEORGIOU et al. "FAILURE-SENSITIVE ANALYSIS OF PARALLEL ALGORITHMS WITH CONTROLLED MEMORY ACCESS CONCURRENCY" Parallel Processing Letters, 2007 [online] [retrieved on 09 October 2015 (09.10.2015)] Retrieved from the Internet <URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.78.2425 >, entire document, especially Abstract; page 4-8	1-11
Y	Boyle et al. "Large-Scale Secure Computation" IACR Cryptology ePrint Archive, 31 May, 2014 [online] [retrieved on 09 October 2015 (09.10.2015)] Retrieved from the Internet <URL: https://eprint.iacr.org/2014/404.pdf >, entire document, especially Abstract; page 23-28	1-11
Y	US 2014/0007250 A1 (STEFANOV et al.), 02 January 2014 (02.01.2014), entire document, especially Abstract; Para [0102]-[0103], [0246]-[0249]	3 and 6
Y	Lorch et al. "Toward Practical Private Access to Data Centers via Parallel ORAM" IACR Cryptology ePrint Archive, 2012 [online] [retrieved on 09 October 2015 (09.10.2015)] Retrieved from the Internet <URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.303.9883 >, entire document, especially Abstract; page 2-3 and 5-8	9-11
A	US 2008/0133841 A1 (FINKLER), 05 June 2008 (05.06.2008), entire document	1-11
A	US 2008/0084889 A1 (ARCHER et al.) 10 April 2008 (10.04.2008), entire document.	1-11

☐ Further documents are listed in the continuation of Box C.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

09 October 2015 (09.10.2015)

Date of mailing of the international search report

04 NOV 2015

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents

P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774